Using Bayes Factors to evaluate evidence for no effect: examples from the SIPS project

Article (Accepted Version)

USING BAYES FACTORS TO EVALUATE EVIDENCE FOR NO EFFECT:

EXAMPLES FROM THE SIPS PROJECT

Zoltan Dienes,

School of Psychology,

University of Sussex

Simon Coulton,

Centre for Health Services Studies,

University of Kent

&

Nick Heather,

Department of Psychology,

Northumbria University

Declarations of interest: None.

Correspondence to Professor Zoltan Dienes, School of Psychology, University of Sussex, Falmer, Brighton, BN1 9QH. Email: dienes@sussex.ac.uk; Phone: +44 1273 876638

ABSTRACT

Aims: To illustrate how Bayes Factors are important for determining the effectiveness of interventions.

Method: We consider a case where inappropriate conclusions were publicly drawn based on significance testing, namely the SIPS Project (Screening and Intervention Programme for Sensible drinking), a pragmatic, cluster-randomized controlled trial in each of two healthcare settings and in the criminal justice system. We show how Bayes Factors can disambiguate the non-significant findings from the SIPS Project and thus determine whether the findings represent evidence of absence or absence of evidence. We show how to model the sort of effects that could be expected, and how to check the robustness of the Bayes Factors.

Results: The findings from the three SIPS trials taken individually are largely uninformative but, when data from these trials are combined, there is moderate evidence for a null hypothesis (H0) and thus for a lack of effect of brief intervention compared with simple clinical feedback and an alcohol information leaflet ($B = 0.24$, $p = 0.43$).

Conclusion: Scientists who find non-significant results should suspend judgment – unless they calculate a Bayes Factor to indicate either that there is evidence for a null hypothesis (H0) over a (well-justified) alternative hypothesis (H1), or else that more data are needed.

KEYWORDS: Non-significance, Bayes Factors, Evidence of absence, Alcohol brief interventions, SIPS Project

Introduction

The dominant approach to statistical inference in randomised controlled trials (RCTs) in addiction and related sciences is null hypothesis significance testing (NHST). However, when no significant differences on outcome measures between intervention and control groups are found, NHST is crucially uninformative [1I. it is unable to distinguish between two interpretations of non-significant findings: (i) there is no evidence that the population means of the groups differ (absence of evidence), or (ii) there is evidence that the population means do not differ (evidence of absence) [2]. The present article provides an illustration of this use of Bayes Factors to address this problem.

When Fisher introduced significance testing he made clear that one should suspend judgment if findings are non-significant [3, 4]. Yet instances of using non-significance to assert the null hypothesis are still frequently found in the literature [5]. This cannot be regarded as an arcane or trivial matter. The SSRI paroxetine was originally said to carry no increased risk of suicide in children on the basis of a non-significant result but was later found to contain such a risk [6:59-62].

There are two unfortunate consequences of the inability of NHST to demonstrate evidence of absence. First, where absence of evidence is concluded from non-significant findings, there may nevertheless have been good evidence that the postulated effect did not exist if the data had been evaluated in an informative way. Thus, where a conclusion may be warranted, the data and information available are wasted. Secondly, when evidence of absence is incorrectly concluded under NHST, there may nevertheless be a real effect of intervention in the population and, in this situation, an opportunity to support an effective intervention by further research will have been missed. Both these kinds of negative consequence will have had retarding effects on theory, research and practice (a problem not fully addressed by power [5]).

A solution to this problem would be a method of statistical inference that gave an actual degree of evidence for the alternative versus the null hypothesis. Such a method is provided by Bayes Factors [7, 8]. While under NHST, only two conclusions are possible from the results of an RCT, either the null hypothesis is rejected or it is not, from a Bayesian perspective there are three: (i) there is

sufficient evidence for the alternative hypothesis that, for example, an intervention has an effect on participants' behaviour; (ii) there is sufficient evidence for the null hypothesis that the intervention has no effect over the alternative considered; (iii) the data are insensitive in distinguishing the hypotheses.

To determine which of these conclusions applies to any given data-set, one calculates the Bayes Factor (B). This is the ratio of how well the observed data are predicted by the alternative hypothesis over how well they are predicted by the null hypothesis. If this ratio is greater than 1, the alternative hypothesis is to that degree supported over the null hypothesis; if it is less than 1, the null hypothesis is supported over the alternative; and if it is about 1, neither hypothesis is supported more than the other. To arrive at a decision in practice, recommended cut-offs [1,9] are that B greater than 3 represents 'substantial' [9] (or better 'moderate' [10]) evidence for the alternative hypothesis and B less than 1/3 represents 'substantial' (or 'moderate') evidence for the null hypothesis, with values in between representing a range of weak evidence for either hypothesis depending on whether B is greater or less than 1. A B of 3 has been shown to correspond roughly with a p-value of 0.05 in conventional statistical testing [1].

The SIPS Project

This project consisted of a pragmatic, cluster RCT in each of two healthcare settings, primary health care (PHC) and accident and emergency services (A&E), and a similar trial in the criminal justice system (CJS). Each trial had a 'step-up' design involving three groups in which components were successively added: (i) a control group consisting of the provision of a Patient Information Leaflet (PIL) together with brief feedback of screening results (i.e., whether or not the patient was drinking at a hazardous or harmful level); (ii) a brief advice group consisting of 5 minutes of structured advice about drinking plus the PIL; (iii) a brief lifestyle counselling group consisting of 20 minutes of counselling preceded by brief advice and followed by the PIL, and given to those patients who returned for a subsequent consultation following the brief advice session. The hypotheses tested were that both interventions

would result in greater reductions in hazardous or harmful drinking than simple clinical feedback of screening results and alcohol information provided by the PIL.

The primary outcome measure in all three trials was whether or not the score on the *Alcohol Use Disorders Identification Test* (AUDIT) [11] was above the cut-point for a designation of hazardous drinking. The main analysis was by intention-to-treat but there was also a *per protocol* analysis which included only those patients who had received a complete intervention and were successfully followed up. For further details of all three trials, including large sample sizes and high follow-up rates, see the corresponding protocol [12-14] and outcome papers [15-17].

In each trial all three groups showed reductions in the proportion of participants classified as hazardous drinkers or worse by the AUDIT but there were no statistically significant differences between groups on this measure at either 6-month or 12-month follow-up in any of the three trials [2-4]. This applied to both intention-to-treat and *per protocol* analyses. Neither were there significant differences between groups on any other alcohol outcome measure (i.e., mean AUDIT score or extent of alcohol problems).

Heather [18] discussed several ways in which the SIPS findings had been misunderstood and the potential effects on research and practice. A prime example of 'proving the null hypothesis' appeared in an article [19] in the magazine *Pulse*, which is widely read by GPs and other health professionals. The article began "GPs should give patients with problem drinking a leaflet rather than advise them to reduce their alcohol intake," because "The SIPS (PHC) study found informing patients of their drinking levels and offering a leaflet … was just as effective as giving patients five- or 10-minutes of lifestyle counselling".

Aims

The aims of the analysis reported here were: (i) to calculate Bayes Factors in order to disambiguate non-significant findings from the SIPS Project and thus determine whether they represent evidence of absence or absence of evidence; (ii) to illustrate how Bayes Factors can clarify non-significant findings.

Method

A Bayes factor requires a summary of the data (a measure of effect size such as a mean difference or an odds ratio, plus the standard error of that estimate) and in addition a specification of the size of effect predicted. One cannot tell how sensitive a study was to detect an effect without an idea of the size of effect that is possible [4]. To be more precise, the relative evidence for H1 versus H0 provided by the data depends on how well the data are predicted by H1 versus how well the data are predicted by H0 [20]. To know how well the data are predicted by H1, we need a model of H1: a specification of what (range of) effect sizes are plausible according to H1 [21].

In a meta-analytic review of the effect of brief interventions on the proportion of hazardous drinkers (i.e. the same dependent variable as used in SIPS), Ballesteros *et al*. [22] found that brief interventions outperformed minimal interventions and usual care after 6-12 months with OR = 1.55 (with a 95% confidence interval of [1.2, 1.90]). In a previous meta-analysis of brief interventions for alcohol use disorders using a wider range of dependent variables, Moyer *et al*. [23] found an average effect size of d = 0.2 for drinking-related outcomes, including alcohol consumption, for a brief intervention versus a control at six months (with a confidence interval from about 0.1 to 0.3, depending on the exact outcome). Assuming an underlying continuous variable given a binary cut-off, this can be converted to Ln OR = $d \times \pi/\sqrt{3}$ [24] or a Ln OR of 0.36 in this case, which is an OR of 1.4. A standardized effect size is a measure of signal relative to noise; and the amount of noise in a study depends on the details of the experimental design and the precise dependent variable used. Translating effect sizes between different dependent variables should therefore be done cautiously. However, the fact that the two meta-analyses above produce very similar estimates (OR = 1.4 vs 1.55) is reassuring. Thus, we will take the estimate from the meta-analysis [22] based on the same dependent variable as we analyse here, an OR of 1.55, i.e. ln OR = 0.44, as a plausible scale of effect for our brief interventions vs the PIL.

We have estimated a rough scale of the effect size but effects smaller or bigger than this are also plausible. Published effects often overestimate the true population effect size, as indicated by

systematic direct replication attempts in psychology and other disciplines [25-27]. One simple way of modelling this assumption is with a bell shape distribution, for example a normal distribution, centred on zero. Because we are making predictions in one direction, one half of the distribution can be removed, to leave, for example, a half-normal distribution. For a half-normal there is one parameter left to decide, its standard deviation. We scale using the rough effect size already derived, namely SD = 0.44.

If a normal or half-normal distribution is used, the scale factor implies a rough maximum, namely twice the standard deviation, as there is only 5% probability of an effect larger than that. Another distribution often used in Bayes Factors is a Cauchy [8]; the rough maximum implied by a Cauchy is 7 times its scale factor (with 5% of its area beyond that). So the final decision to make is what defines a just plausible maximum more adequately, 2*0.44 or 7*0.44 (corresponding to ORs of 2.4 and 21.8 respectively (see [28]). Given that the confidence interval on the effect sizes reviewed in Ballesteros et al [22] and Moyer *et al*. [23] (and see Wilk *et al*. [29]) were less than twice the mean estimated effect, the half-normal is more appropriate than the half-Cauchy.

In sum, we use a half-normal model of H1 because it respects two general assumptions: i] that smaller effects are more likely than larger ones; and ii] that twice the estimated scale of effect is a rough plausible maximum. We report a check on the robustness of our assumptions below. We notate using a half-normal with an SD of 0.44 to model H1 as $B_{H(0,\ 0.44)}$, the H to indicate a half-normal, the 0 indicates its mode is 0 and the 0.44 indicates the standard deviation [for notation see [1]}. (If there were several theories we could use a Bayes factor to test each against H0 separately.)

Results

Table 1 shows the results for the three trials at six months follow-up, the primary end-point. While the Bayes factors for the individual trials on their own largely indicate insensitive evidence, the effect of the data as a whole needs to be evaluated by combining the mean estimates of the ln OR. This can be done using, for example, the calculator from Dienes [7]

http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm[1]) for obtaining a posterior distribution[2,3], which provides the results in the final two rows.

TABLE 1

Odds ratios for a negative AUDIT result after 6 months (primary outcome)

	BA vs PIL			BLC vs PIL		
	Ln OR [95% CI]	SE	$B_{H(0, 0.44)}$	Ln OR [95% CI]	SE	$B_{H(0, 0.44)}$
PHC	-0.16 [-0.65, 0.33]	0.25	0.34	-0.25 [-0.73, 0.22]	0.24	0.26
A&E	.10 [-1.11, 1.31]	0.62	0.88	-0.37 [-1.16, 1.60]	0.70	0.69
CJS	-0.22 [-0.94, 0.48]	0.36	0.45	-0.31 [-1.08, 0.43]	0.39	0.44
PHC/A&E	-0.12 [-0.58, 0.33]	0.23	0.33	-0.26 [-0.71, 0.18]	0.23	0.24
All data	-0.15 [-0.53, 0.23]	0.19	0.24	-0.27 [-0.66, 0.11]	0.20	0.19

BA = brief advice; BLC = brief lifestyle counselling; PIL = Patient information leaflet

Note: The odds ratios and their 95% confidence intervals are reported in the original papers. Natural logs of these ratios, and of their 95% CI limits, are given in the table. (We take the 95% confidence intervals to be approximations of corresponding credibility intervals with vague priors.) Dividing the width of each 95% CI by 2*1.96 gives the standard error of the estimate. For example, in the top left corner, the standard error for BA vs PIL for PHC is (0.33 - -0.65)/3.92 = 0.25. The Bayes Factor can then be determined using the Dienes [7] online calculator, entering the Ln OR as the "sample mean",

[1] Archived here: http://www.webcitation.org/6s4eJTOTs

[2] The calculator assumes all data are estimates of the same population effect. H0 postulates a fixed effect (namely no effect). Thus, for hypothesis testing, a fixed effects analysis typically serves adequately, because if H0 can be rejected on that basis it can be rejected. For estimation, by contrast, it makes sense to take into account the uncertainty in whether the different trials were drawn from the same population, a possibility not a priori denied by a tested hypothesis.

[3] The posterior distribution represents how probable different population effect sizes are, in the light of data.

the standard error as the "sample standard error"; specifying the plausibility of different population values given H1 is not a Uniform; entering "0" for the mean of the normal, "1" for number of tails (the 0 and 1 are the settings for a half-normal) and then 0.44 for the SD (the half-normal parameter that needs setting according to context). See [1, 28, 33] for further examples of how to use the calculator. This site https://medstats.github.io/bayesfactor.html (archived here: http://www.webcitation.org/6s4oBl7iG) allows a greater range of distributions to be used.

The effect of a brief intervention may be specific to its setting of implementation. For example, the motivation for changing drinking is likely to differ between medical patients and offenders. The Kaner team's two Cochrane reviews of the effects of alcohol brief intervention [30, 31] combine primary care and A&E on the ground that, under a wider definition, both are considered forms of primary health care. The difference between brief advice vs PIL between the PHC and A&E trials was 0.26 (SE = 0.67), $B_{N(0, 0.44)}$ = 0.86[4]; that is, there is no evidence one way or the other for a difference between these studies. The estimates are combined in the penultimate row of Table 1. Similarly, for brief lifestyle counselling vs PIL, the difference between the PHC and A&E trials was 0.46 (SE = 0.74), $B_{N(0, 0.44)}$ = 0.90, again no evidence one way or the other for a difference between the two studies. The penultimate row in Table 1 indicates the importance of meta-analytically combining data. While each study alone did not lead to clear conclusions, overall there is moderate evidence for no effect of the brief interventions compared to control (i.e. Bayes factor less than about 1/3), assuming the sort of effect sizes that have been found before for such brief interventions.

Finally, one may wonder how strong the evidence is when all studies are combined. There was no evidence for a difference between CJS and the combined PHC/A&E for brief advice vs PIL, .10 (SE = 0.43), $B_{N(0, 0.44)}$ = 0.71; and no evidence for a difference for brief lifestyle counselling vs PIL, 0.16 (SE = .50), $B_{N(0, 0.44)}$ = 0.77. Nonetheless, the final row In Table 1 indicates moderate evidence for no

[4] H1 is modelled as a Normal here, hence the "N" in the notation. This allows the theoretical difference between conditions to go in either direction.

effect of the brief interventions compared to control for the combined data and for both comparisons, assuming the sort of effect sizes that have been found before for such brief interventions.

The conclusion is relative to the model of H1 used. The robustness of the conclusion can be checked by determining if the same conclusion follows for different ways of representing the same scientific judgements. We used the review by Ballesteros and colleagues [22] for informing the use of a half-normal distribution; that is, this paper played a key role in our scientific judgments. We might also have used the posterior distribution of the effect size presented by Ballesteros *et al.* as our model of H1 (e.g. [32]). Ballesteros *et al*. estimated the ln OR as Ln(1.55) with a standard error of 0.21. Thus, using the Dienes [7] calculator, H1 could be modelled by a normal distribution with a mean of 0.44 and a standard deviation of 0.21. For BA vs PIL, $B_{N(0.44,0.21)} = 0.10$, and for BLC vs PIL, $B_{N(0.44,0.21)} = 0.09$. In both cases the conclusion remains the same. Thus, the difference in representations of H1 did not substantially change the conclusion. We may also perform a sensitivity analysis with the half-normal presentation to determine how conclusions depends on its standard deviation. Any adjustment to allow the effect to be plausibly larger than we have represented will increase support for H0, leading to the same conclusion. Conversely, if one had good reason, independent of the current data, to doubt the effect could be as large as we have represented, that information could render the evidence insensitive. Specifically, with a half-normal, if the estimated effect was as low as 0.24, the Bayes factor for BLC vs PIL is 0.33. When using 0.24 as the SD of a half-normal, the maximum plausible Ln (OR) is about twice the SD, 0.48 (i.e. the maximum plausible OR is 1.6). That is, the evidence becomes relatively insensitive if an effect above about 0.48 can be ruled out. But based on the meta-analytic reviews we have referred to, current scientific judgment is that the effect could well be larger than this, so our conclusions stand. Further investigation into, for example, missing studies, the role of the significance filter, or the role of analytic flexibility in each study may refine the meta-analyses, and thus change our best scientific judgements about effects. The Bayes Factor stands as a provisional judgment given the current state of evidence.

For reference, Table 2 presents the same statistics for the same dependent variable at 12 months. A similar pattern emerges.

TABLE 2

Odds ratios for a negative AUDIT result after 12 months

	BA vs PIL			BLC vs PIL		
	Ln OR [95% CI]	SE	$B_{H(0, 0.44)}$	Ln OR [95% CI]	SE	$B_{H(0, 0.44)}$
PHC	0.09 [-0.63, 0.44]	0.27	0.67	-0.01 [-0.51, 0.48]	0.25	0.48
A&E	0.00 [-0.92, 0.92]	0.47	0.73	-0.11 [-1.20, 0.96]	0.55	0.71
CJS	0.10 [-0.63, 0.85]	0.38	0.77	-0.36 [-1.08, 0.39]	0.38	0.40
PHC/A&E	0.07 [-0.39, .53]	0.23	0.58	-0.03 [-0.47, 0.42]	0.23	0.43
All data	0.08 [-0.31, 0.47]	0.20	0.57	-0.11 [-0.50, 0.27]	0.20	0.29

Discussion

We show that non-significant results from a family of trials involving a large number of participants do not provide compelling evidence for no effect, despite the way the trial results have in fact been interpreted to justify claims of the lack of effectiveness of the interventions. For individual published trials, the results were scarcely moderate evidence for the null hypothesis over a plausible alternative. A problem arises from taking non-significance to mean evidence for H0. Even a high-powered non-significant result does not necessarily mean there is evidence for the null hypothesis over the alternative [28]. However, we also showed that taking a key two trials, or else all three, in the family of trials together provided moderate evidence for no effect of brief advice or brief lifestyle counselling over simple clinical feedback and an alcohol information leaflet at 6 months follow-up compared to

the size of effect one might expect given meta-analyses of brief interventions. There was also moderate evidence for no effect of brief lifestyle counselling at 12-months follow-up. Thus, combining data is useful not just for establishing there is an effect, but in establishing there is not. These conclusions did not follow from the mere fact of obtaining non-significant results; they had to carefully argued. Further, for each study taken separately the evidence is not yet there one way or the other; this conclusion itself depends on using Bayes factors and does not follow from p values alone.

The journal *Addiction* recommends that authors do not report no difference for non-significant findings unless a Bayes Factor has been calculated [2]. In a recent analysis of papers in *Addiction*, only 20% of non-significant findings in a sample of randomised controlled trials were evidence of no effect [33].

A Bayes factor compares how well one model predicts the data (e.g. H1) compared to another model (e.g. H0). The axioms of probability show this is exactly the amount by which one should change one's degree of belief in H1 over H0. Given that evidence is defined as the amount by which one should change belief, the Bayes factor is a measure of strength of evidence [20]. And given that the models are adequate approximations of the scientific theories they represent, there is thus the guarantee from the axioms of probability that the corresponding Bayes factor is a measure of the strength of evidence for the theory of, for example, an effect of a brief intervention over the theory that the intervention is ineffective.

To model H1, we followed the recommendation of Dienes and McLatchie [28] of using a half-normal with the standard deviation set by prior research, in our case by a relevant meta-analysis. The model is based on simple assumptions and the results of other studies; in that sense, it is objective. But further evidence concerning what the theory should predict may revise the outcome.

We modelled H0 as the point prediction of absolutely no difference. It might be objected that this H0 must be wrong; there is never no difference between groups. But the question is whether the model is a good enough approximation, not that it is absolutely true. For example, the assumption of a normal distribution is never absolutely true for any real study; but the issue is whether or not it is a

good enough approximation. Instead of a point null hypothesis we could use an interval null hypothesis [1, 34]. Would an odds ratio between 0.95 and 1.05 be so small as to indicate that the intervention was not worth the cost? If so, instead of a point null, we can set a uniform distribution over the null region [ln (0.95), ln(1.05)] as the model of H0. With this null region hypothesis, the above Bayes Factors remain the same to within .01. That is, the point null hypothesis was a good enough approximation.

Further discussion of the implications of this analysis for research and practice of alcohol brief interventions may be found in Appendix 1 in a supplementary file online. Scientific inference always depends on extra-statistical considerations, such as what control groups are appropriate, and the relation of statistical models to theory [5].

Conclusions

We recommend that scientists who find non-significant results should suspend judgment – unless they calculate a Bayes Factor to indicate either that there is evidence for H0 over a (well justified) H1, or that indeed more data are needed [33, 35]. The present analysis suggests that the findings from the 3 SIPS trials taken individually are largely uninformative but that, when data from these trials are combined, there is moderate evidence for a lack of effect of brief intervention in the SIPS trials. We recommend Bayes factor to be used as standard in clinical trials - and indeed in all hypothesis tests[5].

Declaration of interests

SC and NH were Principal Investigators on the SIPS project but have no other potential conflicts of interest to declare. ZD has no conflicts of interest to declare.

Acknowledgements

[5] The Dienes calculator can be used on the output of any statistics program that gives a parameter estimate and its standard error, so long as the normal approximation is used for hypothesis testing (i.e. if the program uses a *z* or *t* test for significance testing).

We are grateful to Eileen Kaner and Colin Drummond for useful comments on an earlier draft of this article.

References

1. Dienes Z. Using Bayes to get the most out of non-significant results. *Front Psychology* 2014: **5**; 781.

2. West R. Using Bayesian analysis for hypothesis testing (Editorial). *Addiction* 2015; **111**: 3-4.

3. Fisher R. *The Design of Experiments*. London UK: Macmillan; 1971.

4. Johansson, T. (2011). Hail the impossible: *p*-values, evidence, and likelihood. *Scandinavian Journal of Psychology 52*, 113–125.

5. Dienes Z. How Bayes Factors change scientific practice. *J Math Psychol* 2016; **72:** 78-89.

6. Goldacre B. *Bad Pharma.* London UK: Fourth Estate; 2012.

7. Dienes Z. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke UK: Palgrave Macmillan; 2008.

8. Rouder J., Speckman P., Sun D., Morey R., Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bull Rev* 2009; **16**: 225-237.

9. Jeffreys H. *The Theory of Probability.* Oxford UK: The Clarendon Press; 1939.

10. Lee M.D., Wagenmakers E-J. *Bayesian Cognitive Modeling: A Practical Course.* Cambridge UK: Cambridge University Press; 2013.

11. Saunders J. B., Aasland O. G., Amundsen A., Grant M. Alcohol consumption and related problems among primary health care patients: WHO Collaborative Project on early detection of persons with harmful alcohol consumption. I. *Addiction* 1993; **88:** 349-362.

12. Kaner E., Bland M., Cassidy P., *et al*. Screening and brief interventions for hazardous and harmful alcohol use in primary care: a cluster randomised controlled trial protocol. *BMC Public Health* 2009; **9**: 287.

13. Coulton S., Perryman K., Bland M*., et al*. Screening and brief interventions for hazardous alcohol use in accident and emergency departments: a randomised controlled trial protocol. *BMC Health Services Research* 2009; **9**: 114.

14. Newbury-Birch D., Bland M., Cassidy P*., et al*. Screening and brief interventions for hazardous and harmful alcohol use in probation services: a cluster randomised controlled trial protocol. *BMC Public Health* 2009; **9**: 418.

15. Kaner E., Bland M., Cassidy P., *et al*. Effectiveness of screening and brief alcohol intervention in primary care (SIPS trial): pragmatic cluster randomised controlled trial. *BMJ* 2013: **346**: e8501.

16. Drummond C., DeLuca P., Coulton S., *et al*. The effectiveness of alcohol screening and brief intervention in emergency departments: a multicentre pragmatic cluster randomized controlled trial. *PLoS One* 2014; **9**: e99463.

17. Newbury-Birch D., Coulton S., Bland M., *et al*. Alcohol screening and brief interventions for offenders in the probation setting (SIPS Trial): a pragmatic multicentre cluster randomized controlled trial. *Alcohol Alcohol* 2014; **49**: 540-548.

18. Heather N. Interpreting null findings from trials of alcohol brief interventions. *Front Psychiatry* 2014; **5**: 85.

19. Woodhead D. Patient leaflet enough to tackle problem drinking, researchers suggest. *Pulse*; 11 January, 2013. http://www.pulsetoday.co.uk/clinical/therapy-areas/addiction/patient-leaflet-enough-to-tackle-problem-drinking-researchers-suggest/20001448.article

20. Morey, R.D., Romelin, J.-W., Rouder, J.N. The philosophy of Bayes factors and the quantification of statistical evidence. *J Math Psychol* 2016; **72**: 6-18.

21. Lee, M. D., & Vanpaemel, W. (2016). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 1-14.

22. Ballesteros, J., Duffy, J.C., Querejeta, I., Arino, J. Gonzalez-Pinto, A. (2004). Efficacy of brief interventions for hazarous drinkers in primary care|: systematic review and meta-analysis. *Alcohol Clin Exp Res* 2004; **28:** 608-618.

23. Moyer A., Finney J.W., Swearingen C.E. ,Vergun P. Brief interventions for alcohol problems: a meta-analytic review of controlled investigations in treatment-seeking and non-treatment-seeking populations. Addiction 2002; **97:** 279-292.

24. Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. *Introduction to Meta-Analysis*. Chichester UK: Wiley; 2009.

25. Camerer, C.F., Dreber, A., Forskell, *et al.* Evaluating replicability of laboratory experiments in economics. *Science* 2016; **351**: 1433-1436.

26. Ioannides J.P.A. Why most discovered true associations are inflated. *Epidemiology* 2008; **19:** 640-648.

27. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015; **349**: 943–951.

28. Dienes, Z., & McLatchie, N. Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin & Review* 2017;. DOI 10.3758/s13423-017-1266-z

29. Wilk, A. I., Jensen, N. M., & Havighurst, T. C. Meta-analysis of randomized control trials addressing brief interventions in heavy alcohol drinkers. *J Gen Intern Med 1997;* **12**: 274-283.

30. Kaner, E.F.S., Beyer, F., Dickinson, H.O., *et al.* Effectiveness of brief alcohol interventions in primary care populations. *Cochrane Database of Systematic Reviews*2007 , Issue 2. Art. No.: CD004148. DOI:10.1002/14651858.CD004148.pub3.

31. Kaner, E. Screening and brief alcohol intervention in primary care: a perfect fit or a square peg in a round hole? Keynote presentation at Annual Conference of *International Network on Brief Interventions for Alcohol & Other Drugs* (INEBRIA), Lausanne, Switzerland, 22 September 2016. http://inebria.net/wp-content/uploads/2016/10/Plenary-session-1-Eileen-Kaner-SBI-in-PHC-perfect-fit-or-round-peg-in-a-square-hole-final.pdf Archived at: http://www.webcitation.org/6s4gKLyew

32 Verhagen, J., & Wagenmakers, E. J. Bayesian tests to quantify the result of a replication attempt. *J Exp Psychol Gen* 2014; **143**: 1457-1475.

33. Beard E., Dienes Z., Muirhead C., West R. Using Bayes factors for testing hypotheses about intervention effectiveness in addiction research. *Addiction* 2016; **111**: 2230-2247.

34. Morey, R. D., and Rouder J. N. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods* 2011*:* **16**: 406–419.

35. Schönbrodt, F., & Wagenmakers, E.-J. Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review* In press.